STAT 211: Final Exam
Prof. Dan Gillen
University of California, Irvine

Quantifying Breast Cancer Risk Associated with Age at First Birth, Obesity, and Menopause.

Abstract
Data on approximately 800000 women aged 35-84 were analyzed to determine the association of age at first child birth and obesity with the odds of breast cancer development within one year from mammography. Women who gave birth before age 30 were found to have 23% lower (aOR 0.7710; 95% CI 0.7085, 0.8391; $P < 0.0001$) odds of breast cancer development relative to women with no children or who have their first birth after 30. This effect was found to be the same for premenopausal and postmenopausal women. Postmenopausal women who have BMI ≥ 35 were found to have 20% greater (aOR 1.202; 95% CI 1.099, 1.314; $P < 0.0001$) odds of breast cancer development relative to postmenopausal women with BMI in the range of 10-24.99. This was the only patient subpopulation considered for which BMI was significantly associated with odds of breast cancer development. A model predictive of breast cancer is also developed, with a test set AUC of 0.634.

Contents

Introduction

Despite recent decreases in breast cancer incidence rates, it remains the most prevalent cancer among women in the U.S. Previous studies have identified a number of important factors related to the risk of breast cancer development, but the connections between many of these factors are still poorly understood. Moreover, accurate prediction of breast cancer occurrence from clinical factors remains a difficult problem. The further identification and quantification of the role of important risk factors may allow physicians to develop a more thorough understanding of breast cancer risk in subpopulations of women, and perhaps lead to new treatment options and lifestyle recommendations.

From previous research results, it is believed that hormonal factors play a significant role in breast cancer risk. In particular, it is hypothesized that the hormonal changes associated with child birth may have a preventative effect, but the magnitude of this effect is not clear, and it is not known if the preventative benefits from child birth change as a women ages.

Obesity is also known to be significantly associated with increased risk of many forms of cancer, and previous results have found links between BMI, hormone replacement therapy, and postmenopausal breast cancer risk [1]. However, due to the relative rarity of breast cancer cases, studies have been limited in the clear identification of risk in some patient subpopulations by insufficiently large sample sizes. It is likely that there are many relations between these risk factors themselves, for instance, BMI has been shown to be associated with age of menopause [2]. Thus, large studies on new data

and validation analyses are still of great use in understanding the mechanisms of breast cancer risk.

The present study seeks to determine relative brest cancer risk associated with a woman's age at first child birth and obesity, and assess whether the association between these factors and breast cancer risk differs among premenopausal and postmenopausal women. Furthermore, we develop a model predictive of breast cancer development within one year of mammography based on clinical covariates.

1. Materials and Methods

1.1 Patients

A total of N = 2392998 patient records were collected from seven mammography registries from the Breast Cancer Surveillance Consortium. The patients consist of women aged 35-84 who had a screening, diagnostic, or self-reported mammogram, excluding women with previous breast cancer or breast augmentation. Ten clinical variables are recorded in the data set, along with the breast cancer outcome for each patient (occurence within one year from the mammography). These are described briefly in Table 8. Among these covariates, our predictors of interest are age at first birth and BMI, the latter of which we take as a proxy measure for obesity.

1.2 Analysis

The present analysis is concerned with three main objectives:

1. Quantify the association of breast cancer risk and age at first birth, and determine if menopause status modifies this association.

2. Quantify the association of breast cancer risk and BMI, and determine if menopause status modifies this association.

3. Develop a predictive model for breast cancer occurrence.

For (1) and (2) we constructed separate binomial regression models relating age at first birth and BMI to breast cancer risk. The models include the respective main effects, interaction with menopaus and the predictor of interest, and adjustment covariates as described below. Test for significance of the interaction terms and the predictors of interest was conducted with the likelihood ratio test. The significance level for all tests conducted was taken to be 0.05, no adjustments for multiple comparisons were made.

To build a predictive model for breast cancer, we first split all complete cases into a training set (70% of observations) and a validation set. From the training data a binary regression model was chosen by stepwise AIC, considering all covariate main effects and possible interaction terms for inclusion in the model. With this model we then applied 10-fold cross-validation to determine an optimal cutoff (measured by overall accuracy) for prediction of breast cancer. The predictive accuracy of the model was estimated by application to the validation set, and overall performance of the model in the cross-validated training set and test set was summarized with ROC curves.

2. Results

2.1 Descriptive Statistics

All covariates used in this analysis are categorical. The absolute number and proportion of subjects for each covariate value are given in Table 1. Barplots of the covariates across levels are provided in Figures 3 & 2. We note that in the original data description for BMI, the level ranges given as 10-24.99, 25-29.99, and $\geq$ 35, leaving the range 30-34.99 unaccounted for. The last level should likely be $\geq$ 30, but we retained the original level descriptions for consistency with the data source. Additionally, we recode the HRT covariate into a true indicator, so that premenopausal and postmenopausal women who are not using HRT are both coded as 0, and postmenopausal women on HRT are coded as 1.

There are a large number of missing or unknown covariate values, which is of particular concern for the predictors of interest: age at first child (agefirst) and BMI. Of the N = 2392998 total subjects, there are a total of 630453 complete cases (not including missing values for surgical menopause). Over half the data set has missing information for age at first child and BMI (55.5% and 55.8% missing respectively). The large amount of missing data is likely due to the extended time period and geographic regions over which the data was collected. The structure of the missingness across the covariates is illustrated in Figures 4 & 5. From the prior figure, we see that BMI and age at first birth form the majority of missing values, and they tend to be missing together. The latter figure further shows that observations with missing values for age at first birth tend to have missing values for menopause, BMI, number of relatives with breast cancer, previous breast procedures, and HRT use. This correlation of missing is somewhat helpful in that we lose fewer observations when we restrict ourselves to complete cases than we might in the case of unrelated missing values across covariates, but it also raises concerns regarding potential biases introduced into the analysis by non-ignorable missingness. The complete cases do appear to be somewhat representative of the larger sample with respect to the total incidence rate of cancer (0.49% compared to 0.50% incidence rates respectively). In the subsequent models, we used all complete cases across the included covariates. For the model concerned with age at first birth, 800213 distinct observations were used. For the BMI model, 779478 distinct observations were used. This choice omits a substantial portion of the data set, but we are still left with a very large number of observations for use in modeling.

2.2 Inferential Analysis

2.2.1 Modeling Breast Cancer Risk by Age at First Birth

To analyze the risk of breast cancer associated with a patient's age at first birth, a binomial regression model was constructed, adjusting for the following covariates: race, age group, current HRT use, number of cases among 1st degree relatives, previous breast procedures. Tace and age group are included as potential confounders with age at first birth. The additional covariates were included as precision variables since these covariates have been shown to be risk factors for breast cancer, and can reasonably be assumed to be weakly correlated with a woman's age at first birth. Menopause status was included as a main effect and effect modifier on age at first birth. Density was not included in the model, as a change in breast density due to childbirth may be part of the mechanism of action of the protective effect of childbirth, if one exists.

The results of fitting the main effects model (no menopause interaction) are provided in Table 3, including 95% robust confidence intervals of odds ratio estimates for each covariate. The model estimates that women who were under 30 years of age when they first gave birth have 33% lower odds of developing breast cancer relative to women with no children and who are similar across the other covariate values (aOR 0.7710; 95% CI 0.7085, 0.8391; $P < 0.0001$). The odds of developing breast cancer for women who were over 30 years old when they first gave birth are estimated to not be significantly different than the odds for women with no children, assuming similarity across the adjustment covariates (aOR 0.9833; 95% CI 0.8807, 1.0978; $P = 0.764$). Thus we can conclude that there is indeed significantly lower odds of developing breast cancer associated with child birth, but that this benefit is only significant for women who first give birth before age 30.

Considering next the hypothesis that the effect of child birth on the relative odds of developing breast cancer may differ in premenopausal and postmenopausal women, we tested the significance of this effect modification using the likelihood ratio test of the above main effects model with the same model with the effect modifier included (Table 6). The test gave $P = 0.9545$, thus there is insufficient evidence to conclude that there is a significant difference in the association of age of first birth with the adjusted relative odds of breast cancer for premenopausal women compared to postmenopausal women. That is, the lower relative odds for women who had their first child before 30 years of age is roughly the same regardless of menopause status.

High leverage and influential observations for the age first main effects model are given in Table 9 and illustrated in Figure 7. The high leverage points do not appear to have excessive influence, and their covariate values do not appear to be erroneous.

Table 1. Selected Patient Characteristics

Covariate	N (%)
Breast Cancer Cases	11638 (0.0049)
Menopause	
Pre-menopausal	568215 (0.2375)
Post-menopausal	1642824 (0.6865)
Unknown	181959 (0.076)
Age Group	
35-44	330039 (0.1379)
45-54	815558 (0.3408)
55-64	597653 (0.2498)
65-74	435010 (0.1818)
75-84	214738 (0.0897)
Density	
Nearly all fat	148209 (0.0619)
Scattered fibroglandulars	782384 (0.3269)
Heterogeneously dense	674008 (0.2817)
Extremely dense	136011 (0.0568)
Unknown	652386 (0.2726)
Race	
White	138015 (0.7263)
Asian/Pacific Islander	102998 (0.043)
African American	121534 (0.0508)
Other/Mixed	50647 (0.0212)
Unknown	379804 (0.1587)
BMI	
10-24.99	508897 (0.2127)
25-29.99	325352 (0.1360)
≥ 35	222644 (0.093)
Unknown	1336105 (0.5583)
Age at First Birth	
< 30	722195 (0.3018)
≥ 30	141287 (0.0590)
Nulliparous (no children)	201222 (0.0841)
Unknown	1328294 (0.5551)
Num. of 1st Deg. Relative Cases	
0	1718360 (0.7181)
1	295768 (0.1236)
≥ 2	15551 (0.0065)
Unknown	363319 (0.1518)
Previous Breast Proc.	
Yes	420430 (0.1757)
No	1722256 (0.7197)
Unknown	250312 (0.1046)
Surgical Menopause	
Natural	717966 (0.3000)
Surgical	427332 (0.1786)
Pre-menopausal	568215 (0.2374)
Unknown	679485 (0.2839)
Current HRT	
Yes	729196 (0.3047)
No	683350 (0.2856)
Pre-menopausal	568215 (0.2374)
Unknown	412237 (0.1723)

ID 48203, which includes 19 observations, had substantially higher influence relative to the other points, perhaps due to corresponding to women in the youngest age group who have had previous breast procedures, which is a relatively rare group comprising about 2% of the total sample, and only .17% of the subset used for the age first model. This highlights a substantial danger in omitting incomplete observations from the model, and caution is advised in generalizing the conclusions of this analysis without further study of the missingness mechanisms in the data.

2.2.2 Modeling Breast Cancer Risk by BMI

Similar to the age first model, a binomial regression model was constructed to assess the association of BMI and odds of breast cancer. The BMI main effects model includes all adjustment covariates from the previous model, along with the BMI covariate. Age at first birth was excluded from this model. This model found only BMI$\geq$ 35 to be significantly associated with higher odds of developing breast cancer relative to the referent group of BMI in the range 10-24.99. For women in the highest BMI category, the odds of developing breast cancer were estimated to be 14% high relative to the referent group, assuming similar values across the adjustment variables (aOR 1.1399; 95% CI 1.0369, 1.253; $P = 0.0067$). From this model there was insufficient evidence to conclude that the odds of breast cancer for women with BMI in the range 25-29.99 are different than the odds in the referent group (aOR 1.036; 95% CI 0.9517, 1.127; P0.418). These results are summarized in Table 4.

Adding the menopause-BMI effect modifier to this model produced the estimates given in Table 5. With respect to BMI, postmenopausal women with BMI $\geq$ 35 were found to have significantly different odds of breast cancer, estimated to be 20% higher than the referent group (aOR 1.202; 95% CI 1.099, 1.314; $P < 0.0001$). The odds ratios for the other combinations of menopause status and BMI group were found to be not significantly different from the referent groups. Thus BMI appears to significantly increase the odds of breast cancer only for postmenopausal women in the highest BMI category. Comparing the above main effects and interaction models for BMI with the likelihood ratio test showed that the interaction term is significant, $P = 0.0193$ (Table 7).

Influential observations for the BMI interaction model are listed in Table 10 and illustrated in Figure 8. The data point with ID 259521 is somewhat worrisome, showing high leverage and much higher Cook's distance relative to the other data points. This data point corresponds to a single obsevation, a postmenopausal woman in the 45-54 age group, of Other/Mixed race, with BMI in 10-24.99, and on HRT, and who did not develop cancer. We do not apply any remedial action regarding this observation, as it consists of only one observation among nearly 800000 observations used in the model, and so should not have excessive influence on the model coefficient estimates.

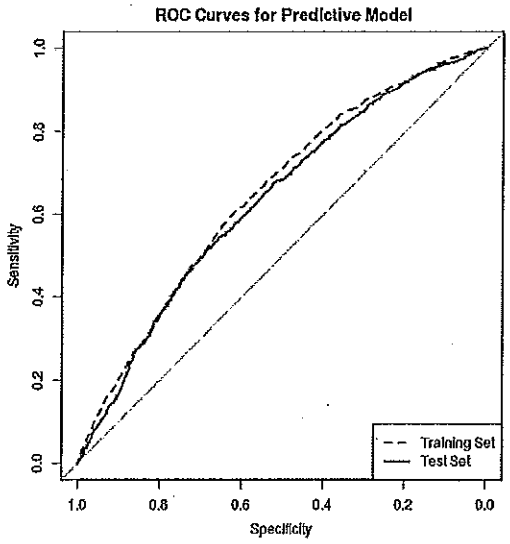

Figure 1. ROC curves for predictive model applied to training and test sets. The AUCs are 0.649 and 0.634 respectively, indicating that the model does somewhat better than random assignment.

Regarding both regression models produced, it is reasonable to assume that there is no correlation of odds of breast cancer among the patients conditional on the included covariates. Thus the variance structure imposed by the binary regression model is correct.

2.3 Predictive Modeling

For construction and assessment of the predictive model we considered only complete observations in order to consider all data set covariates for inclusion in the model. From the complete case data, we first divided the observations into a training set consisting of approximately 70% of the observations, with the remaining observations placed in the test set for later validation of the predictive model.

The covariates included in the model were selected using AIC stepwise selection, considering all main effects and all possible interaction terms. The selected main effects were: age group, density, race, BMI, age at first birth, number of 1st degree relatives with breast cancer, previous breast procedures, HRT use; interaction terms selected were: age group-density, BMI-HRT, and race-HRT.

For a given subject i with covariate structure $\vec{x}_i$, the model produces a linear predictor $\eta_i = \vec{x}_i^T \vec{\beta}$. A prediction of whether subject i will develop breast cancer can then be made according to a cutoff value c such that the model predicts that i will develop breast cancer if $\eta_i \geq c$. Different choices of c produce different combinations of specificity and sensitivity.

A comparison of ROC curves for the model applied to the

training set and test set is given in Figure 1. The area under the curves is 0.649 and 0.634 for the training and test sets respectively, indicating that the model does better than random assignment, but is somewhat poor overall at discriminating between patients that will develop cancer and those that will not. The weakness of this predictive model is likely due, in part, to the rarity of breast cancer in the sample.

The observations considered are only the complete cases, which only make up approximately 19% of the data set. The model accuracy can likely be improved by utilizing the incomplete data in some fashion. Using mode imputation for the missing values (i.e. replacing all missing values with the mode of the covariate) and reproducing the model selection steps from above produces the ROC curves given in Figure 6. The resulting model is slightly different, but the area under the curve for the test set is 0.63, very close to that found for the complete case model. A more refined imputation method is likely required to make substantial gains from the observations with missing values.

A variety of other methods exist for constructing predictive models, and it may be possible to achieve much greater accuracy using an alternative method. For example, future analyses could consider the application of random forests or support vector machines, which have shown good predictive accuracy with similar data sets [3].

Table 2. 10-Fold cross-validation estimates of sensitivity and specificity by cutoff value.

Cutoff	Sensitivity	Specificity
-6.72	1.00	0.03
-6.52	0.99	0.06
-6.32	0.98	0.09
-6.12	0.94	0.17
-5.92	0.89	0.26
-5.72	0.85	0.35
-5.52	0.78	0.43
-5.32	0.67	0.55
-5.12	0.53	0.68
-4.92	0.36	0.80
-4.72	0.22	0.89
-4.52	0.11	0.95
-4.32	0.04	0.99
-4.12	0.01	1.00

2.4 Exploratory Analysis

From the previous model fits for age first and BMI considered above, it was found that all of the adjustment covariates were strongly significant except for race group. As expected, the relative odds of developing breast cancer are much greater in older patient subpopulations. From the BMI interaction model, women in the age group 75-84 have an estimated 3.9 times the odds of developing breast cancer relative to women aged 35-44, assuming similarity across the adjustment covariates (aOR 3.91; 95% CI 3.127, 4.88; $P < 0.0001$). Both the age first main effects model (Table 3) and BMI interaction model (Table 5) gave similar estimates for the effects of age group, race, number of cases among 1st degree relatives, previous breast procedures, and current HRT use, indicating that the risk associated with these covariates is mostly independent of the risk associated with age at first birth and BMI.

Comparing other models, beyond the *a priori* models chosen, with the likelihood ratio test revealed the optimal model (by this measure) includes the main effects for age at first birth, BMI, menopause, age group, HRT use, number of relatives with breast cancer, previous breast procedures, and breast density, as well as effect modifiers for BMI with age group, and BMI with HRT use. The coefficient estimates for this model are given in Table 11. Other considered interaction terms that were rejected in favor of this model were terms for HRT with age group, BMI with menopause, and BMI with age at first birth. We also note that race was not included in the model chosen by successive likelihood ratio tests.

3. Discussion

From the above analysis, there is statistically significant evidence of an association between age at first birth and breast cancer risk such that women who have their first child before age 30 are at a 33% decreased risk of developing breast cancer relative to women who have never had children and who are otherwise similar across the adjustment covariates. This reduced risk was not present for women who had their first child after age 30. There is no statistically significanct evidence of a differential association of age at first birth and breast cancer risk comparing premenopausal and postmenopausal women. The age first main effect model adjusted for menopause, age group, and race. The effect of each childbirth then may be attributable to the hormonal changes that are a consequence of pregnancy. It is plausible the protective effect of pregnancy is through changes in the breast tissue directly, and may not be able to protect against genetic damage already accrued in older women. Secondary modeling also found that the age at first birth effect persists when adjusting for breast density, and was found to be included in the optimal model (as measured by the likelihood ratio test), although this is a data driven conclusion and may be spurious.

There is also statistically significant evidence of an association between BMI and breast cancer risk, and this association was found to change between premenopausal and postmenopausal women. Only postmenopausal women with BMI ≥ 35 were found to show significantlty different odds of breast cancer, with an estimated 20% greater odds of development relative to postmenopausal women with BMI in the range of 10-24.99. Considering BMI as a proxy for obesity, the data suggest that obesity is an important risk factor for development of breast cancer in postmenopausal women. Nu-

merous previous studies have linked BMI with increased risk for many cancers, including breast cancer, but the mechanism of action is not entirely clear. BMI has also been linked with later onset of menopause, so there may be some confounding in the significance of the BMI-menopause interaction, i.e., heavier women tend to have later onset of menopause and thus will tend to have higher odds of developing breast cancer (although the BMI interaction model did adjust for age). In the exploratory modeling, the optimal model included an interaction of BMI with age group and an interaction of BMI with HRT use, but did not include the BMI-menopause interaction. This result is entirely data driven and must be used cautiously, but it does suggest that BMI plays an important role in overall cancer risk independently and through its interaction with other factors. An important goal of future research could be to further determine the causal pathways relating to BMI and other factors for cancer risk.

The logistic regression predictive model developed here performed only somewhat better than random assignment, and would not be useful in a clinical setting with an ROC AUC of 0.63. However, the performance of the model does at least suggest that developing a practical prediction model for breast cancer from the present data set may be possible using other techniques. The logistic model here may suffer from the rarity of breast cancer cases, requiring the accurate prediction of probabilities near 0, and other non-linear classification models may perform better. Due to missing values in the data also required using a small subset of the total data in the prediction modeling, and so further modeling attempts may also achieve better performance by utilizing the incomplete obsevations.

The results from this analysis should be broadly generalizable to women in the U.S. aged 35-84, and who have not previously had breast cancer or breast augmentation, but a more thorough analysis of the structure of the missing data is required to validate the above conclusions and the population to which they can be generalized.

4. Tables

Table 3. Main effects model for age at first birth

Covariate	Odds Ratio Est. (Robust 95% CI)	*P*-value
Age at First Birth		
Nulliparous (no children)	Referent	
< 30	0.7710 (0.7085, 0.8391)	< 0.0001
≥ 30	0.9833 (0.8807, 1.0978)	0.7640
Menopause		
Pre-menopausal	Referent	
Post-menopausal	0.8191 (0.7159, 0.9370)	0.0036
Age Group		
35-44	Referent	
45-54	1.6697 (1.4315, 1.9476)	< 0.0001
55-64	2.8468 (2.3568, 3.4386)	< 0.0001
65-74	3.4831 (2.8695, 4.2280)	< 0.0001
75-84	3.8489 (3.1434, 4.7127)	< 0.0001
Race		
White	Referent	
Asian/Pacific Islander	0.9760 (0.8654, 1.1008)	0.6924
African American	1.1391 (0.9553, 1.3583)	0.1470
Other/Mixed	0.9927 (0.7978, 1.2351)	0.9474
Num. of 1st Deg. Relative Cases		
0	Referent	
1	1.4027 (1.2949, 1.5195)	< 0.0001
≥ 2	1.8365 (1.4690, 2.2960)	< 0.0001
Previous Breast Proc.		
No	Referent	
Yes	1.4321 (1.3318, 1.5399)	< 0.0001
Current HRT		
No	Referent	
Yes	1.2853 (1.1868, 1.3919)	< 0.0001

Table 4. BMI main effect model estimates

Covariate	Odds Ratio Est. (Robust 95% CI)	*P*-value
BMI		
10-24.99	Referent	
25-29.99	1.0355 (0.9517, 1.1266)	0.4184
≥ 35	1.1399 (1.037, 1.253)	0.0067

Table 5. Model for BMI and BMI:Menopause interaction

Covariate	Odds Ratio Est. (Robust 95% CI)	*P*-value
BMI:Premeno		
10-24.99	Referent	
25-29.99	0.9342 (0.7501, 1.1636)	0.5436
≥ 35	0.8826 (0.6999, 1.1131)	0.2915
BMI:Postmeno		
10-24.99	Referent	
25-29.99	1.062 (0.98, 1.151)	0.145
≥ 35	1.202 (1.099, 1.314)	< 0.0001
Menopause		
Pre-menopausal	Referent	
Post-menopausal	0.7206 (0.6086, 0.8533)	0.0001
Age Group		
35-44	Referent	
45-54	1.6457 (1.3806, 1.9616)	< 0.0001
55-64	2.6524 (2.1480, 3.2752)	< 0.0001
65-74	3.2806 (2.6526, 4.0574)	< 0.0001
75-84	3.9061 (3.1269, 4.8795)	< 0.0001
Race		
White	Referent	
Asian/Pacific Islander	1.0536 (0.8843, 1.2553)	0.5593
African American	1.0113 (0.7875, 1.2987)	0.9300
Other/Mixed	0.9411 (0.7479, 1.1843)	0.6047
Num. of 1st Deg. Relative Cases		
0	Referent	
1	1.3632 (1.2553, 1.4805)	< 0.0001
≥ 2	(1.7150, 1.3640) 2.1562	< 0.0001
Previous Breast Proc.		
No	Referent	
Yes	1.4290 (1.3242, 1.5420)	< 0.0001
Current HRT		
No	Referent	
Yes	1.2925 (1.1921, 1.4013)	< 0.0001

Table 6. Results of the likelihood ratio test for significance of menopause-agefirst interaction.

	Resid. Df	Resid. Dev	Df	Deviance	pValue
Agefirst Main Effects Model	15798.00	6273.44			
Agefirst Interaction Model	15796.00	6273.35	2	0.093	0.9545

Table 7. Results of the likelihood ratio test for significance of menopause-BMI interaction.

	Resid. Df	Resid. Dev	Df	Deviance	pValue
BMI Main Effects Model	14832.00	5742.35			
BMI Interaction Model	14830.00	5734.45	2	7.9	0.0193

Appendix A: Additional Tables

Table 8. Data set variable descriptions.

Variable Name in Raw Data Set	Description
n.obs	Number of observations corresponding to this covariate structure.
n.cancer	Number of breast cancer cases diagnosed within one year of mammography for the observations in this covariate structure.
menopaus	Menopause status, 3 levels: premenopausal, postmenopausal, unknown.
agegrp	35-44, 45-54, 55-64, 65-74, 75-84.
density	Measure of breast tissue density, 4 levels ranging from "almost all fat" to "extremely dense".
race	Race of the patient, 4 levels: caucasian, Asian/Pacific Islander African American, Other/Mixed.
bmi*	Body mass index (kg/m^2), 3 levels: 10-24.99, 25-29.99, ≥ 35.
agefirst	Patient's age at first birth, 3 levels: $< 30, \geq 30$, or nulliparous.
nrelbc	Number of 1st degree relatives with breast cancer, 3 levels: 0, 1, ≥ 2.
brstproc	Indicator for previous breast procedures for benign diseases.
surgmeno	Inidicator for surgical menopause.
hrt	Indicator for current HRT use, recoded so that premenopausal women are 0.

* Note that there is a potential error in the original data description, with the BMI range of 29.99-34 unaccounted for; the last level should likely be ≥ 30, but this should be verified with the data source.

Table 9. Observations with greatest leverage for age first main effect model.

ID	Num. Obs.	N. Cancer Cases	Menopause	Age Group	Race	Age First Birth	Rel. BC	Breast Proc.	HRT
8241	6	1	Postmeno	45-54	Asian/Pacific Isln.	Nulliparous	0	No	No
20306	1	0	Postmeno	75-84	African American	>= 30	1	No	Yes
48023	19	0	Premeno	35-44	White	>= 30	0	Yes	No
48025	28	0	Premeno	45-54	White	>= 30	0	Yes	No

Table 10. Observations with greatest leverage and influence for BMI interaction model.

ID	Num. Obs.	N. Cancer Cases	Menopause	Age Group	Race	BMI Birth	Rel. BC	Breast Proc.	HRT
12095	1	0	Postmeno	65-74	African American	25-29.99	1	Yes	No
12096	1	0	Postmeno	65-74	African American	25-29.99	1	Yes	No
25952	1	0	Postmeno	45-54	Other/Mixed	10-24.99	1	No	Yes
25954	4	0	Postmeno	65-74	Other/Mixed	10-24.99	1	No	Yes
48287	2	0	Premeno	45-54	Other/Mixed	25-29.99	0	Yes	No
259521	1	0	Postmeno	45-54	Other/Mixed	10-24.99	1	No	Yes

Table 11. Coefficient estimates for data drive model chosen by the likelihood ratio test.

	exp(Est)	ci95.lo	ci95.hi	z value	Pr(>|z|)
(Intercept)	0.00	0.00	0.00	-45.31	0.00
bmi25-29.99	1.37	0.97	1.94	1.78	0.07
bmi>= 35	0.93	0.58	1.49	-0.32	0.75
menopausPostmeno	0.91	0.77	1.06	-1.21	0.23
agegrp45-54	1.44	1.13	1.83	2.94	0.00
agegrp55-64	2.52	1.91	3.34	6.48	0.00
agegrp65-74	2.57	1.92	3.44	6.37	0.00
agegrp75-84	3.11	2.30	4.22	7.32	0.00
hrtYes	1.34	1.17	1.53	4.21	0.00
nrelbc1	1.29	1.17	1.42	5.09	0.00
nrelbc>= 2	1.73	1.34	2.22	4.25	0.00
brstprocYes	1.35	1.24	1.47	6.78	0.00
agefirst< 30	1.02	0.88	1.18	0.27	0.79
agefirst>= 30	0.86	0.76	0.97	-2.44	0.01
densityScattered fibro.	2.38	1.90	2.97	7.66	0.00
densityHetero dense	3.55	2.84	4.44	11.10	0.00
densityExtremely dense	4.30	3.33	5.54	11.24	0.00
bmi25-29.99:agegrp45-54	0.83	0.56	1.23	-0.93	0.35
bmi>= 35:agegrp45-54	1.61	0.97	2.69	1.83	0.07
bmi25-29.99:agegrp55-64	0.85	0.57	1.27	-0.78	0.43
bmi>= 35:agegrp55-64	1.55	0.92	2.61	1.67	0.10
bmi25-29.99:agegrp65-74	1.16	0.77	1.74	0.72	0.47
bmi>= 35:agegrp65-74	2.27	1.34	3.83	3.07	0.00
bmi25-29.99:agegrp75-84	0.92	0.60	1.42	-0.37	0.71
bmi>= 35:agegrp75-84	2.37	1.36	4.13	3.06	0.00
bmi25-29.99:hrtYes	0.82	0.68	1.00	-1.97	0.05
bmi>= 35:hrtYes	0.72	0.58	0.89	-3.03	0.00

Appendix B: Additional Figures

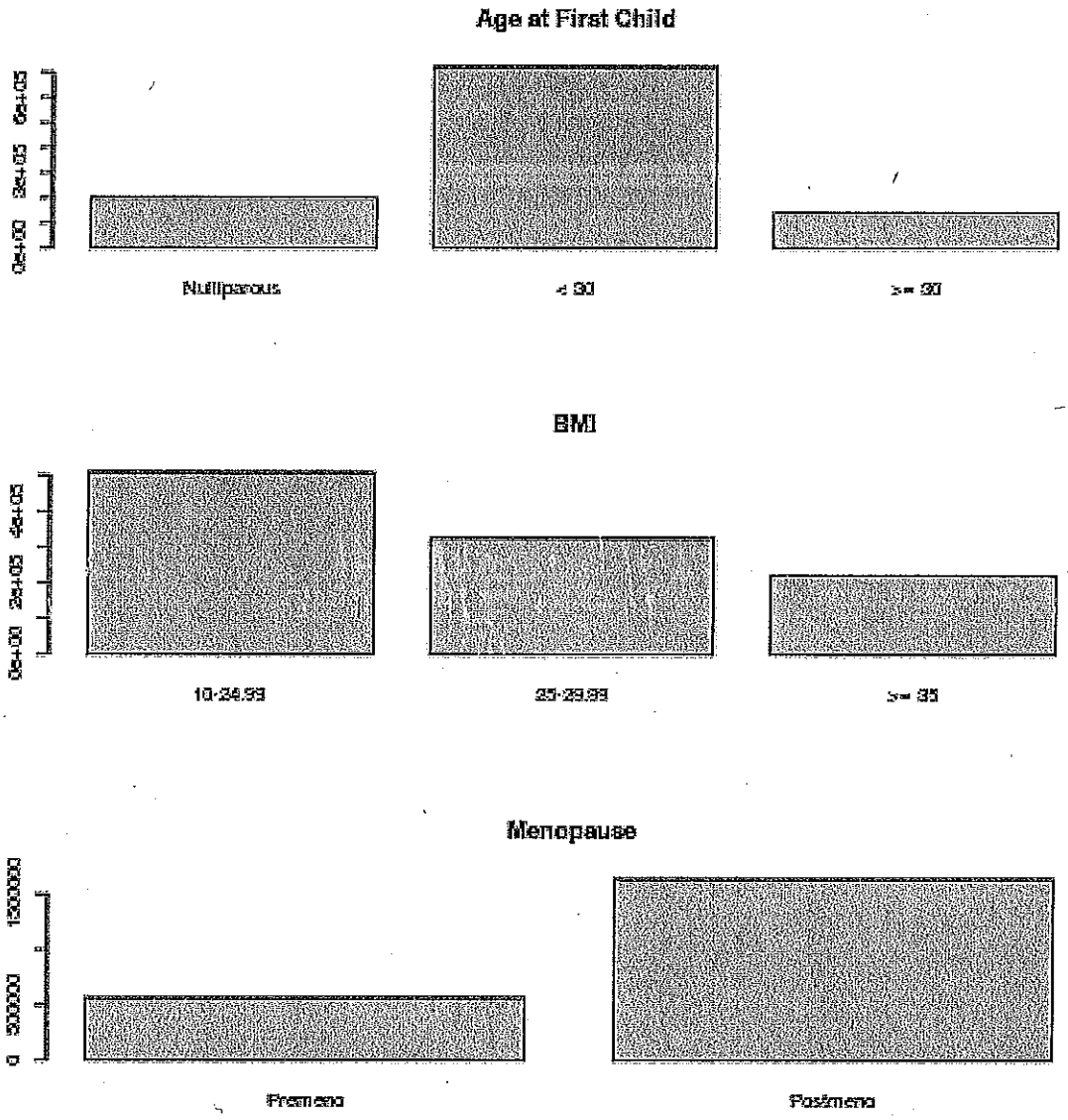

Figure 2. Distrubtion of each of the covariates of interest across levels.

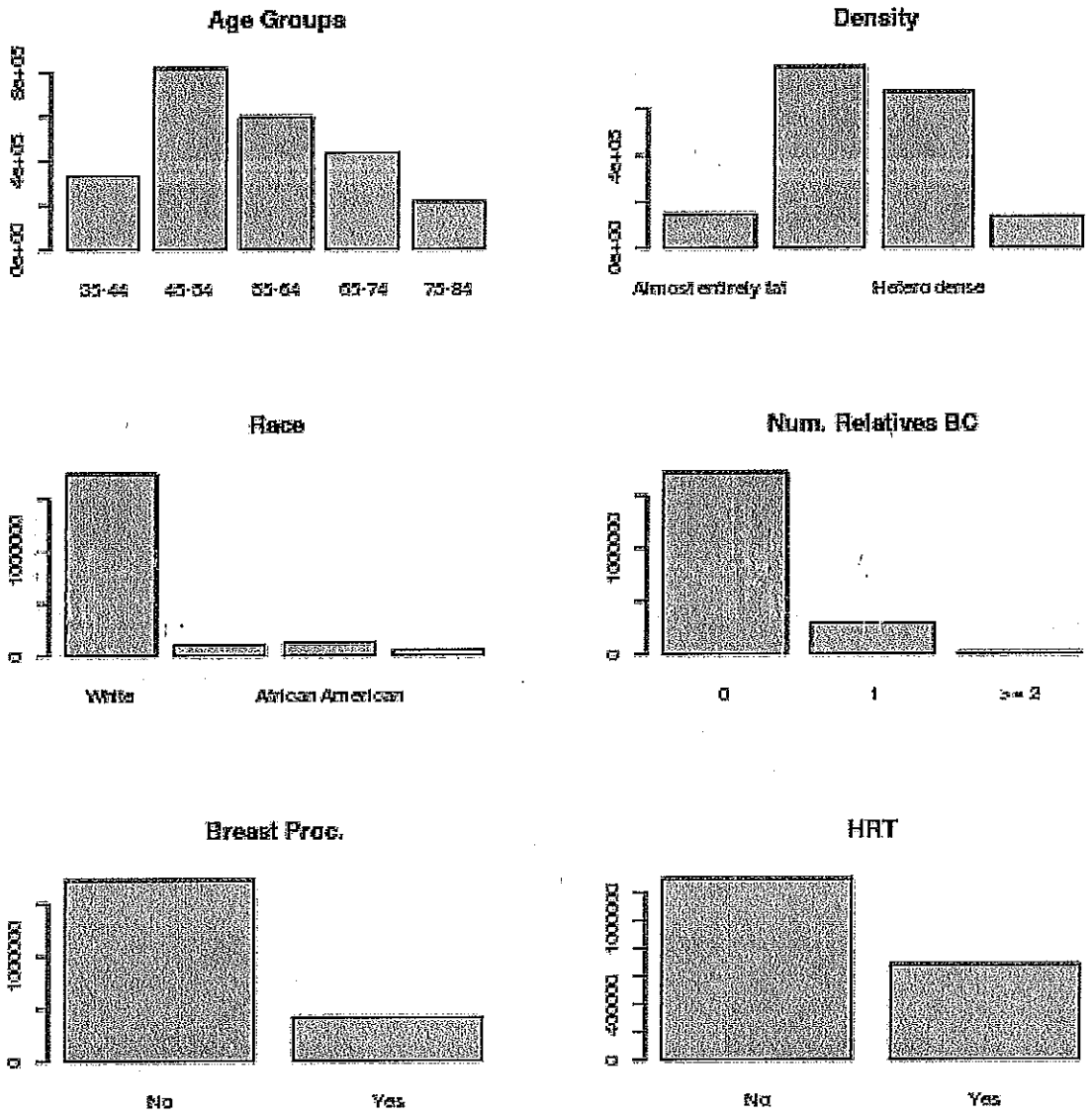

Figure 3. Distribution of adjustment covariates across levels.

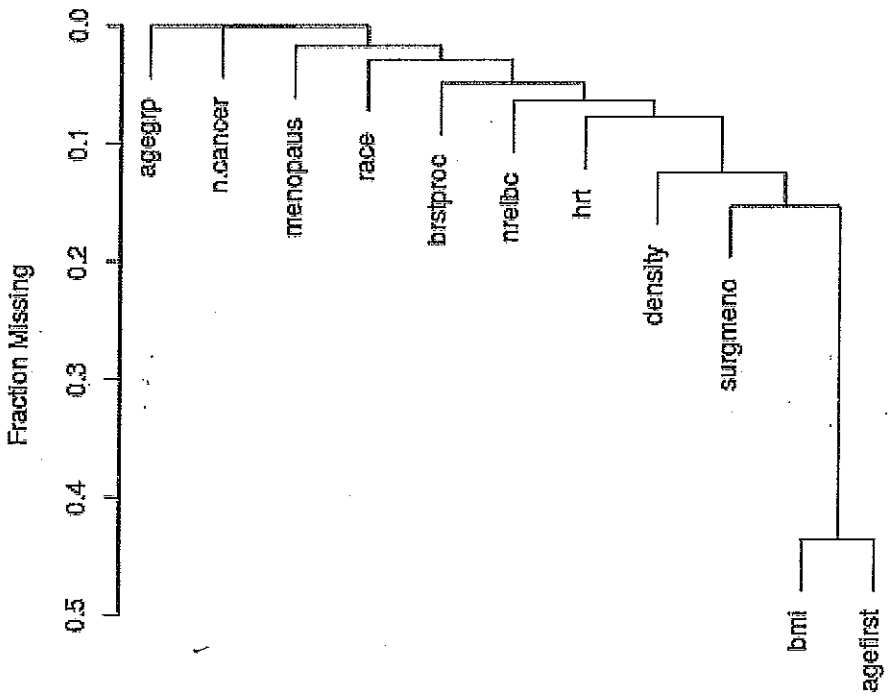

Figure 4. Structure of missing values across covariates. Note that the primary predictors of interest agefirst and BMI are often missing together.

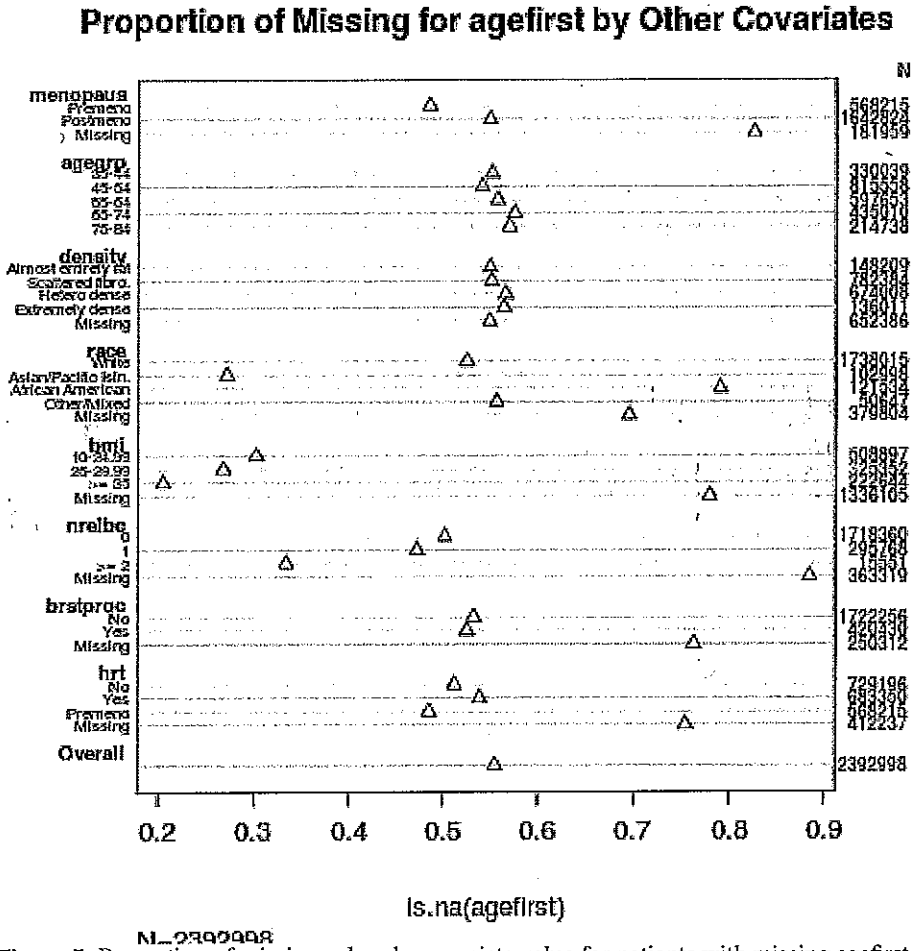

Figure 5. Proportion of missing values by covariate value for patients with missing agefirst.

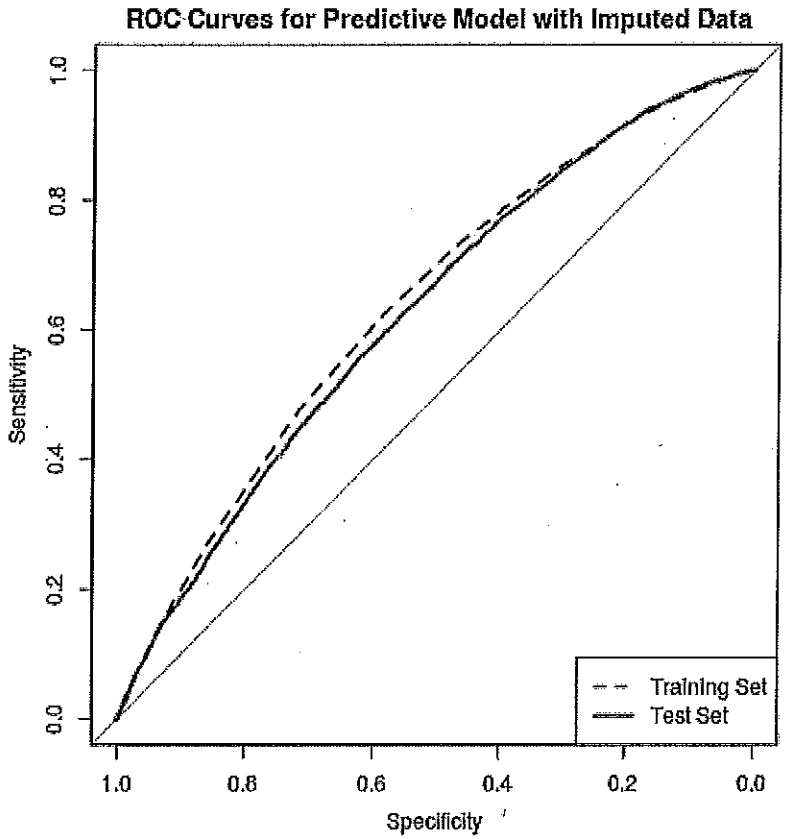

Figure 6. ROC curves for predictive model built from mode imputed data. The predictive performance is very similar to the complete case model.

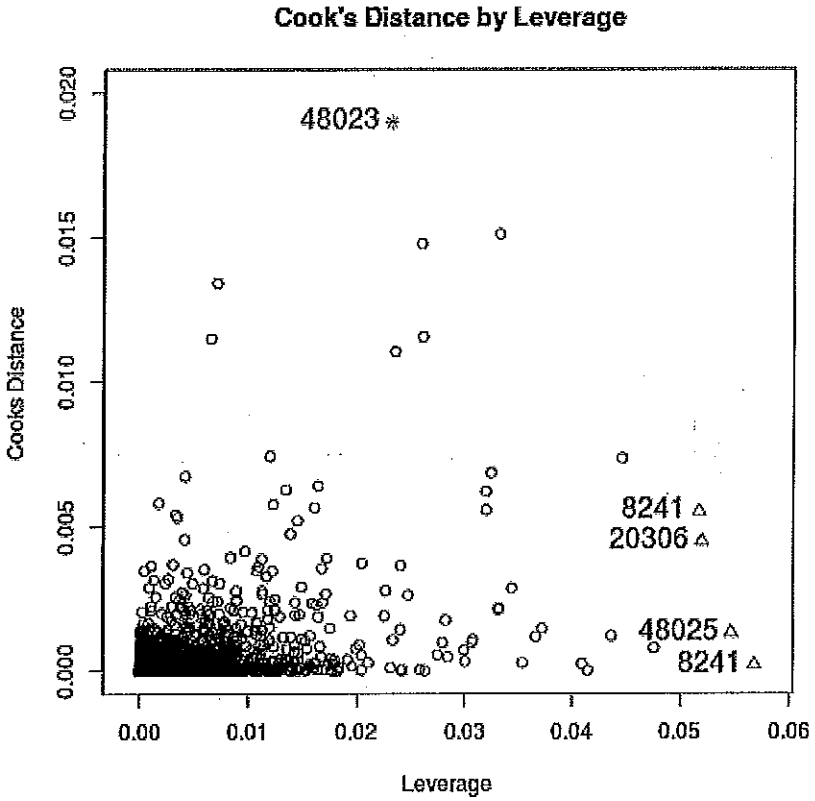

Figure 7. Cook's Distance by leverage plot for age at first birth main effects model. ID numbers for the largest leverage and largest Cook's D are displayed.

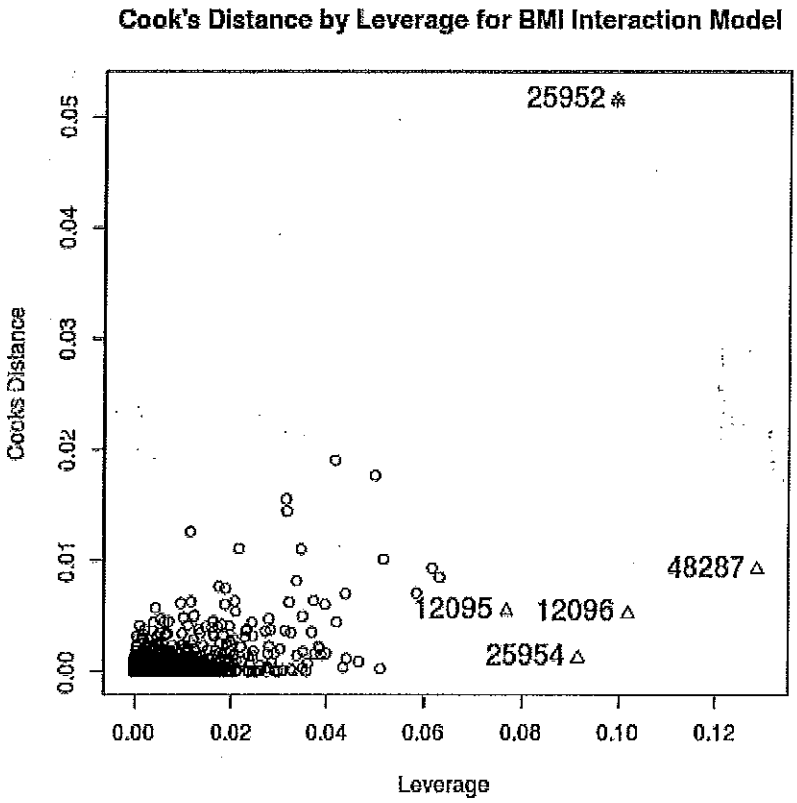

Figure 8. Cook's Distance by leverage plot for BMI interaction model. ID numbers for the largest leverage and largest Cook's D are displayed.

References

[1] Heather Spencer Feigelson, Carolyn R Jonas, Lauren R Teras, Michael J Thun, and Eugenia E Calle. Weight Gain , Body Mass Index , Hormone Replacement Therapy , and Postmenopausal Breast Cancer in a Large Prospective Study. 13(February):220–224, 2004.

[2] M Akahoshi, M Soda, E Nakashima, T Tominaga, S Ichimaru, S Seto, and K Yano. The effects of body mass index on age at menopause. pages 961–968, 2002.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.